Sentiment Analysis and Features Extraction of Reviews using Latent Dirichlet Allocation

Miss. Priyanka V. Chandegaonkar
PG student, Department of Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies, Aurangabad, India.

Mr. K.V.Reddy
Assistant Professor, Department of Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies, Aurangabad, India.

***Abstract*- More recently, the Social Media has become extremely popular to share the user's viewpoints to their friends by using various social networking platforms. It creates obligatory for the users to post their reviews for other users to learn about the quality of the products. But there is much information overloading problems. So to address this problem we propose, A Sentiment Analysis and Features Extraction of Reviews Using Latent Dirichlet Allocation Method used to extract product features, to deal with incorrect sentiment estimation, improve the precision of sentiment mapping, enhance the sentiment dictionary and to calculate the sentiment score and improve prediction accuracy and reliability in the original recommender systems. As the result, it helps to boost the recommendation performance.**

Keyword: Sentiment Analysis, User Sentiment Measurement, Information retrieval, Features Extraction, Sentiment Estimation.

I .INTRODUCTION

Sentiment Analysis is most important task to find out user interest about particular product or item. We have experience of review websites. It gives us the opportunity to share our views for various products that people likes and purchase. On the web or social networking sites sentiment or opinion are increased day by day. To describe any product or item peoples used opinion for product as sentiment. To buy product online peoples sees valuable reviews and then decide what to buy or not. Actually it is field of study used to analyze people's sentiment or opinion. Sentiment Analysis carried out on variety of reviews. That is product reviews, movie reviews, news, and blog. So there is much information overloading problem on review websites. We observe that in many practical cases, it is important to provide numerical scores than binary decisions. Reviews are divided into two groups, positive and negative. However, it is difficult for customers to make a choice when all candidate products reflects positive sentiment or negative sentiment. It is practically very difficult to analyze reviews and extract opinion from huge number of reviews manually. Thus it is difficult to mine or extract valuable information. And the task that how we can extract product features and how to deal with sentiment mapping. In this paper we describe sentiment based approach it deals with features extraction by using LDA [1]. Then we required user sentiment measurement it deals with the SD, ND and SDD Sentiment words, which are used to describe the product features. Besides, we used sentiment dictionary to calculate sentiment score of review. Sentiment score is used in sentiment mapping [2]. LDA automatically extract features of products. It consists of the number of topics and each topic contains a number of words based on their probability. Proposed System is used to deal with incorrect sentiment estimation, improve the precision of sentiment mapping, and enhance the sentiment dictionary and to calculate the sentiment score and improve prediction accuracy and reliability in the original recommender systems.

II .RELATED WORK

In the sentiment based application feature extraction is the most important method. It is used to extract unique features value of the text document. It is used to detect patterns, predict future observation from big data. It is the process of dimensionality reduction.

Reviews-Based Applications

For the task of recommendation there are also many reviews based work. A bag-of-opinions model is used to predict a user's numeric rating in a product review. And by using model we can develop a constrained ridge regression method for learning scores of opinions.

In this reviews based work to solve a new problem, which is aspect identification and rating, together with overall rating prediction in unrated reviews. A LDA-style topic model is useful. This generates ratable aspects over sentiment and associates modifiers with ratings.

Sentiment-Based Applications

Sentiment analysis carried out on three different levels. Review-level, phrase-level and sentence-level. Review-level analysis and sentence-level analysis attempt to classify the sentiment of a whole review to one of the predefined sentiment polarities, including positive, negative and neutral. While phrase-level analysis is used to extract the sentiment polarity of each feature that a user expresses his or her attitude to the particular features of a particular product. Phrase-level sentiment analysis is used for construction of sentiment lexicon.

D. M. Blei, A.Y. Ng, and M. I. Jordan[1], In this paper, describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA method is used for various purposes such as features extraction. It finds the probability of each topic in text corpora. LDA is a Bayesian model. It is used to model the relationship of reviews, topics and words. They deal with the efficient approximate inference techniques based on variation methods and an EM algorithm for empirical bayes parameter estimation. In this paper LDA is also used in text classification in each task of document. And to train the model of fully observed set of users in collaborative filtering.

Yang Gao, Yue Xu [2] In this paper used the field of information filtering to generate users' information needs from a collection of documents. A main assumption for these approaches is that the documents in the collection are all about one topic. In formation Filtering, clustering of documents using LDA, Topics represented by Patterns using MPBTM, Relevant documents given based on user interests. Used to User information needs are generated in terms of multiple topics. Maximum matching of patterns take place.

Milan Gaonkar[3] In this paper a new approach proposed that uses lexicon database AFINN to assign each word in a text a value called valence. It tries to evaluate each opinion bearing word on the basis of its intensity. In this paper is to classify a given tweet/paragraph whether it is of positive [True positive, False positive] or negative [True negative, False negative] sentiment. In this paper a new approach is been proposed that uses lexicon database to assign each word in a text a value called valence. Valence means how the single word is affecting the whole sentences. They used classifiers. A simple technique for constructing classifiers models that is Naive Bayes that assign class labels to problem instances, represented as vectors of feature values. A discriminative classifier SVM (Support Vector Machine) is formally defined by a separating hyper plane.

Sharmistha Dey [4] in this paper using simple methods like frequency based aspect extraction and lexicon based sentiment analysis on data. This method is useful to find scope for improvements of bucketization. Bucketization is the process of determining the labeled data.

Shoushan Li, Sophia Yat Mei Lee [5] in this paper a machine learning approach that is polarity shifting is used. That is information into a document-level sentiment classification system. For binary classifier on polarity shifting first to generate automatically training data feature selection method is require.

Nathan Aston, Timothy Munson [6] in this paper using Modified Balanced Winnow for sentiment analysis on OSNs they present a stream algorithm. It is more practical to analyze data that is coming from social media.

Table 1. Review Summary

Sr. no	Paper name and Authors	Description	Advantages

1	Latent Dirichlet allocation. (D. M. Blei, A.Y. Ng, and M. I. Jordan)	In this paper, describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora.	Efficient approximate inference techniques based on variation methods and an EM algorithm for empirical Bayes parameter estimation.
2	Pattern-based Topics for Document Modelling in Information Filtering. (Yang Gao, Yue Xu)	In formation Filtering, clustering of documents using LDA, Topics represented by Patterns using MPBTM, Relevant documents given based on user interests.	User information needs are generated in terms of multiple topics. Maximum matching of patterns take place.
3	Sentiment Classification using Product Reviews (Milan Gaonkar)	In this paper a new approach proposed that uses lexicon database AFINN to assign each word in a text a value called valence.	It tries to evaluate each opinion bearing word on the basis of its intensity.
4	Aspect Extraction and Sentiment Classification of mobile App using App store reviews.(Sharmistha Dey)	In this paper using simple methods like frequency based aspect extraction and lexicon based sentiment analysis on data.	This method is useful to find scope for improvements of bucketization.
5	Sentiment classification and polarity shifting(Shoushan Li, Sophia Yat Mei Lee)	In this polarity shifting information into a document-level sentiment classification system. It requires a feature selection method for automatically generate the training data for a binary classifier on polarity shifting detection of sentences.	It is able to consistently improve the overall performance across different domains and training data sizes.
6	Sentiment Analysis on the Social Networks Using Stream Algorithm(Nathan Aston, Timothy Munson)	They present a stream algorithm using Modified Balanced Winnow for sentiment analysis on OSNs	It is more realistic to analyze data that is continually coming from new social media posts.

III .SYSTEM ARCHITECTURE

The proposed system has five noteworthy modules. These modules are Input (User Review Sentences), Data Preprocessing, Features Extraction, User Sentiment Measurement, Sentiment Analysis, Sentiment Score. The figure.1 beneath demonstrates a diagrammatic perspective of the proposed system alongside its modules and their stream of communications. The following sections describe more details about our approach.

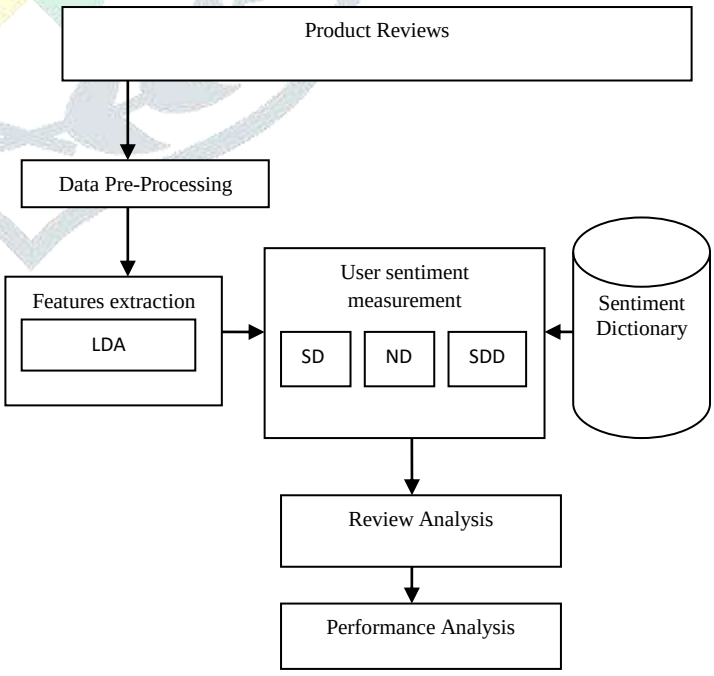


Fig 1. System Architecture

1. Review data

In this first we have used mobile reviews as a collection of data set from Amazon website. It consists of more than 1000 mobile reviews. The mobile corpus comprises real world reviews from website. We first copy all mobile product reviews into csv file. This dataset is derived from the customer's reviews in Amazon Commerce Website for authorship identification. Peoples consider all the valuable reviews of particular product. The major critical factor for the improvement of the quality services rendered and enrichment of deliverable are the user's opinions. Review sites, blogs and micro blogs provide a good understanding of the reception level of products and services. Opinions are the major and actual data or more precise a decision for any user in making a purchase.

2. Data Pre-processing

Pre-processing in our system is a module that involves transforming of raw data into an understandable format. We have copied real world reviews into file so the data is unstructured and noisy. To remove impurities like missing value, smoothing noisy data, Filter out stop words. Stop words are words that are so common in documents that they are useless for analysis. Stop words could be prepositions, pronoun.

3. Extracting Product Features

Extracting Product feature is an important task of review mining and summarization. Product feature extraction is used to find product features that customers refer to in their topic reviews. The main focus of product features is on the discussed issues of a product. Features Extraction is the stages in informational retrieval system, which is used to extract the unique feature values of a text document. To detect patterns, extract information, or predict future observations from big data, informative features are important. It would be useful to characterize the sentiment which the review or express about the products. This model first discovers the targets on which sentiment have been expressed in a sentence, and then determines whether the reviews are positive, negative or neutral. By using LDA we extract product features from textual reviews. We mainly want to get the product features including some named entities and some product or services attributes. LDA (Latent Dirichlet Allocation) is a Bayesian model, which is utilized to model the relationship of reviews, topics and words.

A. Latent Dirichlet Allocation

LDA is topic modeling technique. By using this method we can define documents, No. of topics. LDA is not only useful for extracting meaning of the text , but at the same time it make soft clustering for documents based on topic. LDA automatically perform document clustering. This is an imaginary process.

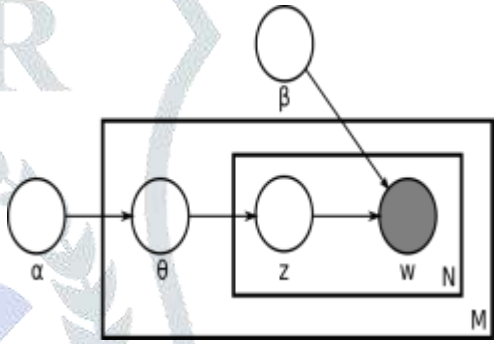
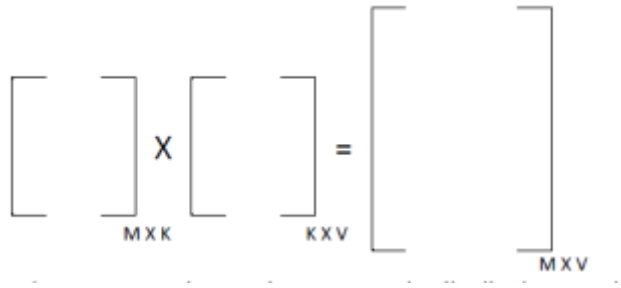


Fig 2. Plate Diagram of LDA

In the above figure larger rectangle denoted by M, which indicates the total no. of documents within the corpus. And smaller rectangle N denotes no. of words in document. All parameters depend on where they fall inside the rectangle. That is at the document level, at the word level or both levels. α is per document topic distribution and β is per topic word distribution. Θ is the topic distribution for document. w is word. z is topic for the n word in document. Input of LDA is collection of preprocess data (Document D). Assign no. of topic determine no. of words in documents. When direct sampling of frequent words is required apply gibbs sampling.



Where,

M X K is per document topic distribution matrix.

K X V is per topic word distribution matrix.

K is the no. of topics.

The Generative process of LDA

The input of LDA model is all users' document sets D, and we assign the number of topic Γ (we set 50 empirically). The output is the topic preference distribution for each user and a topic list, which contains at least 10 feature words under each topic. The generative process of LDA consists of three steps, as follows.

1) For each document d_j, we choose a dimensional Dirichlet random variable θ_m ~Dirichlet (a).

2) For each topic z_k, where k ∈ [1, Γ], we choose φ_k ~Dirichlet (b). For each topic z_k, the inference scheme is based upon the observation that:

$p\ (\Theta,\Phi|Dtrain,\alpha,b) = \sum_z p\ (\Theta,\ \Phi \mid z,\ D^{train},\ \alpha,b) \times P\ (z, \mid D^{train},\ \alpha,b)$.

(3) We obtain an approximate posterior on Θ and Φ by using a Gibbs sampler to compute the sum over z. Repeating the process above and eventually we get the output of LDA.

The TF-IDF method is used to weight a word and shows the importance to that word based on the number of times it appears in the document. That is frequency of word. That is term frequency means number of time term t appears in a document of total number of terms in the documents. And inverse document frequency is total number of documents with the number of documents in term t.

B. Gibbs Sampling

A Gibbs sampler or sampling is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations. The point of Gibbs sampling is that given a multivariate distribution. This is used to determine no. of word in document. First choose a topic mixture for the document over a fixed set of topic. Generate the words in the document by-Pick a topic based on the document multinomial distribution above. Next pick a word based on the topics multinomial distribution. Randomly assign each word in each document to one of the k topics. Suppose we want to obtain k sample of X=(x1,….,xn) from a p(x1, …, xn).

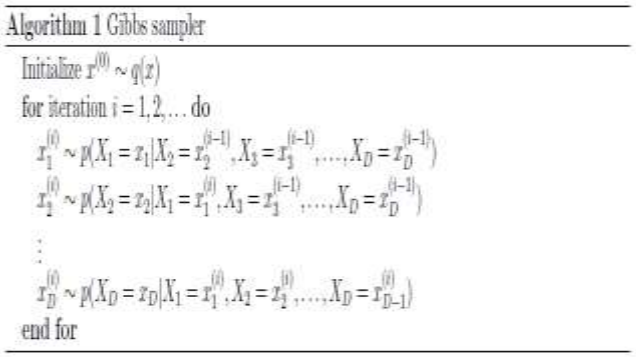


Figure 3. Gibbs Sampling Algorithm

Required distribution of each frequent word first initialize value with 0. In above algorithm Xi is the vector. Repeat the above process k times. Probability of each word is considered.

1. P_1 = P (Topic | document)

-Proportion of words in document that are currently assigned to topic.

2. P_2 = P (word | topic)

- Proportion of assignment to topic over all documents that come from word w.

$P = P_1 * P_2$

4 User Sentiment Measurements

We extend sentiment dictionary to calculate user sentiment measurement. In this we merge positive sentiment word list and positive sentiment evaluations word list of sentiment

dictionary into one list and called it as POS- Words. Also we need to merge the negative sentiment word list and negative evaluation word list of sentiment dictionary into one list and called it as NEG-Words. We have five different levels in sentiment degree dictionary (SDD) which has five different levels of sentiment words.

Table 2. Brief Introduction of the Sentiment Dictionaries

Dictionaries	Representative Words
SD(8938)	POS-Words(4379): attractive, clean, beautiful, comfy, convenient, delicious, delicate, exciting, fresh, happy, homelike, nice, ok, yum ... NEG-Words(4605): annoyed, awful, bad, poor, boring, complain, crowed, dirty, expensive, hostile, sucks, terribly, unfortunate, worse ...
ND(56)	no, nor, not, never, nobody, nothing, none, neither, few, seldom, hardly, haven't, can't, couldn't, don't, didn't, doesn't, isn't, won't, ...
SDD(128)	Level-1 (52): most, best, greatest, absolutely, extremely, highly, excessively, completely, entirely, 100%, highest, sharply, superb ... Level-2 (48): awfully, better, lot, very, much, over, greatly, super, pretty, unusual ... Level-3 (12): even, more, far, so, further, intensely, rather, relatively, slightly more, insanely, comparative. Level-4 (9): a little, a bit, slight, slightly, more or less, relative, some, some what, just. Level-5 (7): less, not very, bit, little, merely, passably, insufficiently.

In the above there is also need the negation dictionary (ND) by collecting frequently used negative prefix words. We have used this sentiment dictionary. The contradiction of something is called negation or denial. Denial of the truth of a clause or sentence, typically involving the use of a negative word or a word or affix with negative force. i.e. no, hardly, never, nothing, not, non, don't, dis, un, in. These words are used to reverse the polarity (polarity shift method is used to detect the polarity shift in the sentence; it helps to improve the performance.) of the sentiment word. (phone is not good) By default value of negation check coefficient is +1.0(not bad).If not then reverse the sentiment polarity and coefficient is set to -1.0 If there is level 1 sentiment degree word before sentiment word, Dw is set a value of 5. If there is level 2 sentiment degree word comes before sentiment word, Dw is set a value of 4 etc. The values of levels are predefined. Dw = [0.25, 0.5, 2, 4, 5]

5. Sentiment Score

We convey positive sentiment by saying "high quality," but "high price" or "high noise" represents the negative sentiment. As a result, such direct rule may result in incorrect sentiment estimation. To improve the precision of sentiment mapping, we use two main linguistic rules as [8] "and" rule and "but" rule. In "and" rule sentences or reviews that are connected with "and"-like conjunctives usually express the same sentiment polarity. For example, "this cup has high quality and nice appearance" indicates that "high" for "quality" and "nice" for "appearance" are of the same polarity. In "and"-like terms consists of: as well as, likewise etc.

"but" rule: sentences or reviews that are connected with "but"-like conjunctives usually indicates the opposite sentiment polarity. For example, "this cup has high price but nice appearance" indicates that "high" for "price" and "nice" for "appearance" are of the opposite polarity. Other "but"-like terms consists of: nevertheless, though, however etc.

To improve sentiment mapping, need to calculate Sentiment Score.

$$S(r) = \frac{1}{N_c} \sum_{c \in r} \sum_{w \in c} Q \cdot D_w \cdot R_w$$

Where,

Nc – no of sentence.

Q – Negation check coefficient.

Dw – set of value.

Rw – initial score of the sentiment word w.

IV. PERFORMANCE EVALUATION

For this implementation we have used Amazon Dataset which we have taken from amazon website. From this amazon dataset we considered input data as a collection of product reviews that is mobile reviews. More than 1000 reviews are considered for result analysis. Results are calculated on the basis of parameters such as Precision, Recall, F-Measure.

Table 3. Result Analysis of different datasets

Dataset	Precision	Recall	F-measures
Movie	**86.3**	**59.2**	**72.7**
SFU	**92.0**	**55.0**	**73.5**

YELP	**91.75**	**60.10**	**87.1**
AMAZON	**89.25**	**86.42**	**90.17**

Table 3 shows result analysis of different datasets. We compared propose system dataset with the previous system datasets that is amazon dataset with the movie, SFU, Yelp datasets. We calculate precision, recall and F-measures of proposed system and compare with the previous system database performance. As a result proposed system gives average result. Figure 4 shows the graphical representation of result of proposed system.

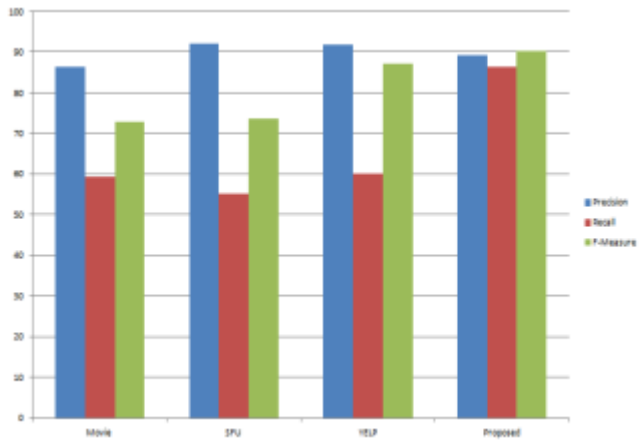


Fig 4. Graphical Representation of Result

V. CONCLUSION AND FUTURE WORK

Basic LDA model can be used to extract product features. This Propose system gives average results. It improved performance. Average result of precision is 89.25, Recall is 86.42, F-Measure is 90.17 as the result, and it helps to boost the recommendation performance. It helps to identify incorrect sentiment estimation. In future by using fine grained sentiment analysis we can enrich the sentiment dictionaries.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan "Latent Dirichlet Allocation", Journal of Machine Learning Research 3 (2003) 993-1022.

[2] Yang Gao, Yue Xu, Yuefeng Li, "Pattern-based Topics for Document Modeling in Information filtering. IEEE Transactions on Knowledge 2015.

[3] Milan Gaonkar, "Sentiment Classification Using Product Reviews," in Proc. Conf. Empirical Methods Natural Language Process., 2011, pp. 1524–1534.

[4] Sharmistha Dey, "Aspect Extraction and Sentiment Classification of Mobile Apps using App-Store Reviews. Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol., 2017, pp. 359–367.

[5] S. Li et al., "Sentiment classification and polarity shifting," in Proc. 23rd Int. Conf. Comput. Linguistics, 2010, pp. 635–643.

[6] Nathan Aston, Timothy Munson, Jacob Liddle, Garrett Hartshaw, Dane Livingston, Wei Hu "Sentiment Analysis on the Social Networks Using Stream Algorithms" in Journal of Data Analysis and Information Processing 2014.

[7] Felipe Bravo-Marquez, Eibe Frank and Bernhard Pfahringer "Positive ,Negative, or Neutral: Learning an Expanded Opinion Lexicon from Emoticon-annotated Tweets," in Proc. AAAI Conf. Artif. Intell., 2012, pp. 1692–1698.

[8] Weishi Zhang and Guiguang Ding, "Generating Virtual Ratings from Chinese Reviews to Augment Online Recommendations", Tsinghua University LI CHEN, Hong Kong Baptist University ACM Trans. Intell. Syst. Technol. 4, 1, Article 9 (January 2013).

[9] L. Qu, G. Ifrim, and G. Weikum, "The bag-of-opinions method for review rating prediction from sparse text patterns," in Proc. 23rd Int. Conf. Comput. Linguistics, 2010, pp. 913–921.